Conditional transformation models for survivor function estimation

Möst, Lisa ; Hothorn, Torsten

Abstract: In survival analysis, the estimation of patient-specific survivor functions that are conditional on a set of patient characteristics is of special interest. In general, knowledge of the conditional survival probabilities of a patient at all relevant time points allows better assessment of the patient's risk than summary statistics, such as median survival time. Nevertheless, standard methods for analysing survival data seldom estimate the survivor function directly. Therefore, we propose the application of conditional transformation models (CTMs) for the estimation of the conditional distribution function of survival times given a set of patient characteristics. We used the inverse probability of censoring weighting approach to account for right-censored observations. Our proposed modelling approach allows the prediction of patient-specific survivor functions. In addition, CTMs constitute a flexible model class that is able to deal with proportional as well as non-proportional hazards. The well-known Cox model is included in the class of CTMs as a special case. We investigated the performance of CTMs in survival data analysis in a simulation that included proportional and non-proportional hazard settings and different scenarios of explanatory variables. Furthermore, we re-analysed the survival times of patients suffering from chronic myelogenous leukaemia and studied the impact of the proportional hazards assumption on previously published results.

DOI: https://doi.org/10.1515/ijb-2014-0006

1 Introduction

The estimation of a patient's individual survival probabilities over time is a key aspect of survival analysis. Technically, we are interested in estimating the conditional survivor function, *i.e.* the probability of surviving up to a specific time point t, conditional on a set of patient-specific explanatory variables. However, common regression models for censored data seldom focus on the direct estimation of the conditional survivor function. Instead, the models concentrate either on the estimation of hazard functions or on summary statistics. In the omnipresent Cox proportional hazards model (Cox, 1972), the conditional hazard function is estimated by cleverly treating the baseline hazard function as a nuisance parameter. Only in a second step can the corresponding conditional survivor functions be derived from this model, for example using the Breslow estimator (*e.g.* Andersen, Christensen, Fauerholdt, and Schlichting, 1983). Hence, if one is interested in the conditional survival probabilities, methods for the direct estimation of the conditional survivor function are required.

Moreover, assumptions associated with common modelling strategies for survival data are restrictive. For example, the Cox model is based on the assumption of proportional hazards, the proportional odds model assumes constant odds ratios over time, and in the parametric accelerated failure time model log-transformed responses imply *e.g.* log-normal-distributed, log-logistic-distributed, etc. survival times. Although remedies are available, such as stratified Cox models or time-varying effects (Sargent, 1997, Xu and O'Quigley, 2000, Scheike and Martinussen, 2004, Tian, Zucker, and Wei, 2005), and although model diagnostics (for example, based on Schoenfeld residuals, or formal misspecification tests Schoenfeld, 1982, Ng'Andu, 1997) and particularly tests for the proportional hazards assumption (for example, based on cumulative sums of martingale-based residuals or weighted residuals Lin, Wei, and Ying, 1993, Grambsch and Therneau, 1994) help to detect unrealistic assumptions, models making less strong assumptions would be widely welcomed.

We suggest estimating the conditional distribution function of the survival times T given a set of patient characteristics x directly in terms of conditional transformation models (CTMs). CTMs have been presented recently in Hothorn, Kneib, and Bühlmann (2014) and allow the direct and semiparametric estimation of the conditional distribution function $\mathbb{P}(T \leq t|X = x)$ under rather weak assumptions. The general model class includes both the proportional odds model and the proportional hazards model as special cases. Nevertheless, the strict assumptions of proportional hazards or proportional odds are relaxed in CTMs. This is achieved by including interaction terms between the survival time and the explanatory variables. For example, the CTM framework allows for varying explanatory variable

effects on the hazard function and hence is able to estimate non-proportional hazards as well. However, this advantage comes at the price of a more complex model, which is not easily communicated by simple parameter estimates or even p-values. Graphic approaches are needed to interpret the model, but we can always fall back on the classical approach when the more-flexible model suggests that it is safe to assume proportional hazards. P-values or confidence intervals cannot be obtained based on large sample thoery, but can be simulated using bootstrap approaches instead. We illustrate model interpretation by means of flexible conditional survivor functions in a re-analysis of a randomised clinical trial comparing busulfan, hydroxyurea, and interferon-α treatment for chronic myelogenous leukaemia. This trial has been analysed earlier using a Cox model (**?**Aalen, 1988, McGilchrist and Aisbett, 1991a, Vaida and Xu, 2000). Since the proportional hazards assumption is questionable for the different treatment groups, we re-analysed the data set using the CTM approach and allowed for non-proportional effects of the patient characteristics over time.

Transformation models play an important role in survival analysis. The one-to-one correspondences between the proportional hazards and the proportional odds model and linear transformation models has already been established in Doksum and Gasko (1990) and Cheng, Wei, and Ying (1995). Cheng, Wei, and Ying (1997) extended the model class to semiparametric transformation models for failure times. Chen, Jin, and Ying (2002) introduced a unified estimation procedure for the analysis of censored data using linear transformation models and Zeng and Lin (2006) proposed a class of semiparametric transformation models to characterise the effects of possibly time-varying covariates on the intensity functions of counting processes. For the estimation of the crude failure probabilities of a competing risk, conditional on explanatory variables, Fine (2001) proposes a semiparametric transformation model. These approaches are based on generalised estimation equations. Our approach uses component-wise gradient boosting methodology for model fitting. This approach has the advantage that it incorporates variable selection and shrinkage of coefficient estimates into the model fitting process. These regularisation techniques for regression models are necessary for the estimation of survival probabilities because patient characteristics are often highly correlated. Hence, prediction accuracy for the survival probabilities can usually be improved if only a subset of the available patient characteristics is incorporated into the prediction formula. Owing to the component-wise fitting procedure, the algorithm can deal with high-dimensional data. Variable selection in high dimensional survival data has also been brought up by Lee, Chakraborty, and Sun (2011) and Van der Vaart and van der Laan (2006). Lu and Li (2008) previously derived a component-wise boosting algorithm for the analysis of survival data in terms of nonlinear transformation models.

Fully nonparametric estimation of the conditional survivor function has also been considered in the past. Making no assumptions about the form of the survivor function can be advantageous over parametric or semiparametric approaches, where the underlying assumptions may be violated. Furthermore, nonparametric approaches can be used to check if one of the more restrictive parametric or semiparametric submodels provides a good fit to the data. The well-known product limit estimator introduced by Kaplan and Meier (1958) enables nonparametric estimation of the unconditional survivor function. Dabrowska (1987), Dabrowska (1989), González Manteiga and Cadarso-Suarez (1994) and Iglesias Pérez and González Manteiga (1999) present generalisations of the product limit estimator by the introduction of kernel-based weights to estimate the *conditional* survivor function nonparametrically. In the light of counting process theory, McKeague and Utikal (1990) propose a general counting process regression model for estimating conditional survivor functions and Li and Doss (1995) propose a class of estimators for the conditional survivor function based on a fully nonparametric model. The usage of local linear estimators for the conditional survivor function is suggested in Spierdijk (2008).

In contrast to kernel-based methods, tree-based approaches and especially random forests can be used to estimate conditional distribution functions rather precisely without relying in strict model assumptions. For right-censored data, Hothorn, Lausen, Benner, and Radespiel-Tröger (2004) introduced a forest aggregation scheme producing estimates of the conditional survivor function. The same scheme was used later by Meinshausen (2006) for uncensored observations; an alternative forest variant (random survival forests) was introduced by Ishwaran, Kogalur, Blackstone, and Lauer (2008). Conditional inference forests (Strobl, Boulesteix, Zeileis, and Hothorn, 2007), based on an aggregation of conditional inference trees (Hothorn, Hornik, and Zeileis, 2006b), use the aggregation scheme introduced by Hothorn et al. (2004) and have been shown to perform akin to other forest variants for right-censored data (Mogensen, Ishwaran, and Gerds, 2012) and were used as a completely nonparametric competitor for conditional transformation models in the empirical experiments here.

Another useful alternative to the Cox model or to linear transformation models is censored quantile regression (e. g. Powell, 1986, Chernozhukov and Hong, 2002, Honoré, Khan, and Powell, 2002, Portnoy, 2003, Peng and Huang, 2008, Wang and Wang, 2009, Wey, Wang, and Rudser, 2014). With this approach, the conditional quantiles of the survival times are modelled in terms of regression models. In contrast to our proposed CTM approach, not all conditional quantiles of the survival times are modelled simultaneously but are instead modelled separately. Hence, quantile crossing (Dette and Volgushev, 2008) is a potential problem of this procedure.

2 Conditional transformation models for survival data

In what follows T denotes a positive random variable describing the time from a well-defined starting point to an event of interest, *e.g.* death or recurrence of a disease. We consider N patients with survival times T_i, $i = 1,\ldots,N$, and a vector of patient characteristics $x_i = (x_{i1},\ldots,x_{ip})$. Since we do not assume that all patients experience the event of interest by the end of the study period and since some patients quit the study early, the event times sometimes are not actual event times but rather right censored. The *observed* right-censored event times $\tilde{T}_i$ are defined by $\tilde{T}_i = \min(T_i, C_i)$, $i = 1,\ldots,N$, where C_i denotes the time under observation or censoring time. Furthermore, the event indicator $\delta_i = I(T_i \leq C_i)$ is 1 for real event times and 0 for right-censored event times. A common assumption is that the survival time T and the vector of explanatory variables X are independent of the censoring time C.

The conditional survivor function S is defined as the conditional probability of being event-free up to some time point t in terms of the conditional distribution function of the survival times given the explanatory variables x:

$$S(t|X = x) = \mathbb{P}(T > t|X = x) = 1 - \mathbb{P}(T \leq t|X = x). \tag{1}$$

When using CTMs, we aim at estimating the conditional distribution function of the survival times via

$$\mathbb{P}(T \leq t|X = x) = F(h(t|x)), \tag{2}$$

and the conditional survivor function can be calculated by the relationship given in Equation 1. Thereby, the conditional distribution function is modelled in terms of the monotone transformation function $h : \mathbb{R} \to \mathbb{R}$, which depends on the patient characteristics x. F denotes an absolute continuous distribution function $F : \mathbb{R} \to [0,1]$ with corresponding quantile function $Q = F^{-1}$. In CTMs, only the monotone transformation function h is estimated, whereas the link function F is chosen a priori.

To embed the well-known class of linear transformation models (Doksum and Gasko, 1990, Cheng et al., 1995) into CTMs exemplarily, we reconsider the formulation of the proportional hazards model in terms of a linear transformation model given in Doksum and Gasko (1990). The conditional distribution function of the survival times resulting from the Cox model can be written as

$$\mathbb{P}(T \leq t|X = x) = F(h_T(t) + x^\top \beta), \tag{3}$$

where F denotes the distribution function of the minimum-extreme value distribution, and the transformation of the survival times $h_T(t)$ equals the logarithm of

the cumulative baseline hazard. In linear transformation models, the influence of the explanatory variables is restricted to linear functions, and most importantly, the transformation function h is decomposed into a part depending only on the survival times $h_T(t)$ and a part depending only on the explanatory variables $x^\top \beta$. This strict decomposition results in the proportional hazards assumption.

In CTMs, the proportional hazards assumption is relaxed by allowing for interactions between the survival times and the explanatory variables in terms of the conditional transformation function $h(t|x)$. Furthermore, we assume additivity on the scale of the transformation function and decompose the monotone transformation function h into J partial transformation functions, whereby each $h_j : \mathbb{R} \to \mathbb{R}$ is conditional on x:

$$\mathbb{P}(T \leq t|X = x) = F(h(t|x)) = F\left(\sum_{j=1}^{J} h_j(t|x)\right). \tag{4}$$

In analogy to the representation of the Cox model in Equation 3, we choose F to be the minimum-extreme value distribution function. In this way, we operate on the same scale of distribution functions in the CTM and the Cox model, and hence estimations from the two approaches are comparable. The CTM given in Equation 4 can be understood as a generalisation of the proportional hazards model to more flexible non-proportional hazard functions, if F is the minimum-extreme value distribution function.

Since all interaction terms between the survival time and the explanatory variables are avoided in the Cox model (Equation 3), the effects of the explanatory variables are estimated to be constant and are not allowed to vary over time. This assumption is relaxed in the more flexible model class of CTMs. Interaction terms between the survival time and the explanatory variables are established in terms of the partial transformation functions h_j that depend on the survival time *and* on the explanatory variables simultaneously (Equation 4). Hence, the explanatory variables effects' are allowed to vary over time, what usually results in non-proportional hazards. We do not only estimate one single parameter for each explanatory variable like it is done in the Cox model. Instead, separate partial transformation functions are defined for each explanatory variable, whereby a smooth parameter function over time is estimated for each group of a categorical explanatory variable. For continuous explanatory variables a smooth parameter surface is estimated, that depends on the survival time and on the continuous explanatory variable.
In comparison to alternative nonparametric approaches, the estimation of the conditional survivor function is no black box procedure in CTMs. Although the model assumptions are weak in CTMs, a certain model structure is imposed by introducing additive partial transformation functions. The resulting explanatory variable

effects' over time can be interpreted and can be illustrated graphically. Hence, concerning model complexity, semiparametric CTMs can be placed inbetween the less flexible semiparametric linear transformation models (e. g. the Cox model) and more flexible nonparametric approaches.
If one is interested in better interpretable versions of CTMs, the model class of conditionally linear transformation models (CTLMs) introduced in Möst, Schmid, Faschingbauer, and Hothorn (2014) can be considered. In CLTMs, the conditional transformation function h is restricted to transformation functions that are linear in the response transformation. Due to this restriction, the explanatory variables are only allowed to influence the conditional mean and the conditional variance of the response transformation, whereas higher moments remain uneffected. The effects of the explanatory variables on the conditional mean and the conditional variance are non-linear, but can be interpreted in CLTMs. Further restrictions of the transformation function are concievable. For example, if all interaction terms between the survival time and the explanatory variables are deleted and the explanatory variable effects' have to be linear, the conditional transformation function of the Cox model (Equation 3) results as a special case. The Cox model can even be further restricted by choosing special forms of the monotone response transformation $h_T(t)$. For example, the specification of $h_T(t) = \log(\lambda) + \nu \cdot \log(t)$ results in the Weibull model.

2.1 Estimating conditional transformation models for survival data

Hothorn et al. (2014) explain thoroughly how CTMs are estimated by the minimisation of the continuous ranked probability score (CRPS) (see Gneiting and Raftery, 2007) using a component-wise boosting algorithm. The CRPS was chosen because it constitutes a proper scoring rule for distributional and probabilistic forecasts (Hothorn et al., 2014). When we estimated CTMs for survival data, we also used a component-wise boosting algorithm to minimise an appropriate integrated loss function. First, we formulated the integrated loss function for uncensored observations and then extended the loss function to right-censored observations.

Integrated loss function for uncensored observations. In an uncensored survival data setup, we observed the survival or event times $T_i, i = 1,\ldots,N$, for N patients under consideration. Furthermore, we considered a grid of time points $\{t_\iota : \iota = 1,\ldots,n\}$ ranging from the study's starting point $t_1 = 0$ to the study's end point t_n. Typical choices for the grid points $\{t_\iota : \iota = 1,\ldots,n\}$ are equally spaced grid points or a grid composed of all distinct survival and event times. Hence, we

were able to observe the binary survival status $I(T_i \leq t_\iota)$ for each patient at each grid point; the status is 1 if the patient experienced the event by t_ι and is otherwise 0.

We aimed at estimating the conditional distribution function of the event times $\mathbb{P}(T \leq t_\iota | X = x) = F(h(t_\iota|x))$ (see Equation 2) in terms of the conditional transformation function h, where t_ι denotes some arbitrary time point in the study period. This estimation problem can be reformulated as estimating the probability $F(h(t_\iota|x))$ for the binary event $T \leq t_\iota$ and is solved by minimising an appropriate loss function. We chose the logarithmic score (or negative binomial log-likelihood) for measuring the loss between the binary event status $T_i \leq t_\iota$ and the corresponding probability $F(h(t_\iota|x_i))$ for N patients at a specific time point t_ι:

$$\begin{aligned} \text{LS}(t_\iota) &= -\frac{1}{N} \sum_{i=1}^{N} \{I(T_i \leq t_\iota) \log(F(h(t_\iota|x_i))) \\ &\qquad + I(T_i > t_\iota) \log(1 - F(h(t_\iota|x_i)))\}. \end{aligned} \tag{5}$$

Alternatively, the Brier score or the absolute error loss can be chosen as an appropriate loss function (Hothorn et al., 2014, Gneiting and Raftery, 2007, Schemper and Henderson, 2000).

Based on the logarithmic score for one specific time point t_ι (see Equation 5), we defined the integrated logarithmic score over all time points, which allows estimation of the whole conditional distribution function $\mathbb{P}(T \leq t | X = x)$ in one step:

$$\begin{aligned} \text{ILS} &= -\frac{1}{N} \sum_{i=1}^{N} \int_0^{t_n} \{I(T_i \leq t) \log(F(h(t|x_i))) \\ &\qquad + I(T_i > t) \log(1 - F(h(t|x_i)))\} \; dW(t), \end{aligned} \tag{6}$$

where $W(t)$ denotes a weight function for the time points. By choosing the same weight $\frac{1}{n}$ for all time points t_ι, $\iota = 1, \ldots, n$, we get the empirical version of Equation 6:

$$\begin{aligned} \widehat{\text{ILS}} &= -\frac{1}{N \cdot n} \sum_{i=1}^{N} \sum_{\iota=1}^{n} \{I(T_i \leq t_\iota) \log(F(h(t_\iota|x_i))) \\ &\qquad + I(T_i > t_\iota) \log(1 - F(h(t_\iota|x_i)))\}, \end{aligned} \tag{7}$$

which is used as the empirical loss function in the boosting algorithm. Of course, other weight functions $W(t)$ for the time points are conceivable.

When the conditional distribution function is estimated, the ultimate goal is to estimate the conditional transformation function h such that the empirical risk in Equation 7 is minimised. The minimisation of the empirical risk is equivalent

to the minimisation of the loss between the true survival status at time point t_ι, $I(T_i \leq t_\iota)$ and the corresponding estimated survival probability $F(\hat{h}(t_\iota|x_i))$ for all time points and all patients. In other words, the survivor function for a specific patient $\hat{S}(t_\iota|x_i) = 1 - F(\hat{h}(t_\iota|x_i))$, $\iota = 1,\ldots,n$, is estimated such that the survival probabilities fit the patient's true survival status best.

Integrated loss function for right-censored observations. In survival analysis, we often face right-censored survival times. Then, we do not observe the true survival time T_i for the right-censored patients, and only the observed survival times $\tilde{T}_i = \min(T_i, C_i)$, $i = 1,\ldots,N$, are available. One way to account for right-censored observations in model estimation is the inverse probability of censoring weighting (IPCW) approach suggested by Van der Laan and Robins (2003) and used often in the past (*e.g.* see Gerds and Schumacher, 2006, Hothorn, Bühlmann, Dudoit, Molinaro, and van der Laan, 2006a). For example, Robins and Finkelstein (2000) present an IPCW version of the Kaplan-Meier estimator and the log-rank test to account for noncompliance and dependent censoring. Van der Laan and Robins (2003) give an IPCW example for right-censored data with time-independent explanatory variables and censoring at random and suggest that the full data loss function (*i.e.* the integrated logarithmic score in our case) be weighted by the inverse probability of censoring weights

$$\omega_{i\iota} = \frac{\Delta(t_\iota)}{\hat{K}(\min(T_i, t_\iota))}, \tag{8}$$

where $\Delta(t_\iota) = I(C_i > \min(T_i, t_\iota))$. $\hat{K}$ denotes the marginal Kaplan-Meier estimator of the censoring distribution, $\hat{K}(t) = \hat{\mathbb{P}}(T > t)$, based on $(\tilde{T}_i, 1 - \delta_i)$, $i = 1,\ldots,N$, hence on the observed survival times and the reverse censoring indicator, which is 1 for right-censored observations and 0 otherwise. Furthermore, the censoring time C_i is set to ∞ for uncensored observations.
To calculate the IPCWs for the integrated logarithmic score in Equation 7 based on Equation 8, we have to distinguish four different situations:

1. Uncensored observations ($\delta_i = 1$) that experience the event up to t_ι ($\tilde{T}_i \leq t_\iota$):

$$\omega_{i\iota} = \frac{I(\tilde{T}_i \leq t_\iota, \delta_i = 1) \cdot \overbrace{I(C_i > T_i)}^{=\Delta(t_\iota)=1}}{\hat{K}(T_i)} = \frac{1}{\hat{K}(T_i)} = \frac{1}{\hat{K}(\tilde{T}_i)}.$$

2. Uncensored observations ($\delta_i = 1$) that do not experience the event up to t_ι ($\tilde{T}_i > t_\iota$):

$$\omega_{i\iota} = \frac{I(\tilde{T}_i > t_\iota, \delta_i = 1) \cdot \overbrace{I(C_i > t_\iota)}^{=\Delta(t_\iota)=1}}{\hat{K}(t_\iota)} = \frac{1}{\hat{K}(t_\iota)}.$$

3. Right-censored observations ($\delta_i = 0$) that experience the censoring up to t_ι ($\tilde{T}_i \leq t_\iota$):

$$\omega_{i\iota} = \frac{I(\tilde{T}_i \leq t_\iota, \delta_i = 0) \cdot \overbrace{I(C_i > T_i)}^{=\Delta(t_\iota)=0}}{\hat{K}(\text{NA})} = 0.$$

4. Right-censored observations ($\delta_i = 0$) that do not experience the censoring up to t_ι ($\tilde{T}_i > t_\iota$):

$$\omega_{i\iota} = \frac{I(\tilde{T}_i > t_\iota, \delta_i = 0) \cdot \overbrace{I(C_i > t_\iota)}^{=\Delta(t_\iota)=1}}{\hat{K}(t_\iota)} = \frac{1}{\hat{K}(t_\iota)}.$$

The resulting weighting scheme corresponds exactly to the weighting scheme given in Graf, Schmoor, Sauerbrei, and Schumacher (1999), which results in a consistent estimator (see Gerds and Schumacher, 2006). In short, the observations are weighted by the inverse probability of not being censored up to the event time (situation 1) or up to the specific time point under consideration (situations 2 and 4). The current survival status is unknown in situation 3; consequently these observations get zero weights. Thus, censored observations contribute to the model estimation process up to their censoring time point and those observations that have already been censored are accounted for in the inverse probability of censoring weights.
We extended the empirical logarithmic score for uncensored observations given in Equation 7 to right-censored observations by including the weighting scheme presented above. Hence, the empirical version of the integrated censored logarithmic score results in

$$\begin{aligned} \widehat{\text{ILS}^C} &= -\frac{1}{N \cdot n} \sum_{i=1}^{N} \sum_{\iota=1}^{n} \Bigg\{ I(\tilde{T}_i \leq t_\iota, \delta_i = 1) \log(F(h(t_\iota | x_i))) \cdot \frac{1}{\hat{K}(\tilde{T}_i)} \\ &\qquad + I(\tilde{T}_i > t_\iota) \log(1 - F(h(t_\iota | x_i))) \cdot \frac{1}{\hat{K}(t_\iota)} \Bigg\}, \end{aligned} \tag{9}$$

which is used as empirical risk function in the boosting algorithm.

2.2 Boosting conditional transformation models for survival data

In CTMs, the conditional distribution function of uncensored responses is estimated using component-wise boosting with penalisation (for a detailed description, see Hothorn et al., 2014). This algorithm has to be slightly modified for the estimation of right-censored survival data. Thereby, the empirical risk given in Equation 9 is minimised with respect to the transformation function h. Furthermore, the parametrisation of the partial transformation functions h_j, $j = 1,\ldots,J$, (Equation 4) has to be slightly adapted for survival data. In component-wise boosting algorithms regularisation is achieved by the application of penalised base-learners. The overall model complexity is regulated by the number of boosting iterations M. For a thorough introduction to component-wise boosting we refer to Bühlmann and Hothorn (2007) and Schmid and Hothorn (2008).

Parametrisation of the partial transformation functions. Considering the parametrisation of the partial transformation functions in Hothorn et al. (2014), we defined for the j-th partial transformation function:

$$h_j(t_\iota|x) = \left(b_j(x)^\top \otimes b_T(t_\iota)^\top\right)\gamma_j, \quad j = 1,\ldots,J, \tag{10}$$

where $b_T : \mathbb{R} \to \mathbb{R}^{K_T}$ denotes the basis along the grid of time points t_ι, $\iota = 1,\ldots,n$, and $b_j : \chi \to \mathbb{R}^{K_j}$ is a basis for (a subset of) the explanatory variables x. Both sets of basis functions are connected via a Kronecker product, whereby an interaction surface between the survival times and the explanatory variables is established. The vector $\gamma_j \in \mathbb{R}^{K_j K_T}$ contains the basis coefficients for the established interaction surface. The basis b_T defines the functional form of the transformation of the survival times, and the functional form of b_j defines how the survival time transformation is influenced by the explanatory variables (Hothorn et al., 2014). Hence, one usually chooses B-spline basis functions for b_T, and depending on the desired flexibility or the measurement level of the explanatory variables, one chooses linear basis functions or B-spline basis functions for b_j. In more detail, linear basis functions are chosen for b_j if x is univariate and categorical or if x is univariate and continuous and a linear influence is assumed. B-spline basis functions are chosen for b_j if x is univariate and continuous and the influence might be more flexible. Additionally, b_j might depend on more than one explanatory variable and appropriate multivariate basis functions have to be considered. The partial transformation functions h_j are typically supposed to be smooth in the first argument t and in the conditioning variable x. This is due to the fact that continuous distribution functions have to be smooth in the response variable. Moreover, we expect similar distribution functions

for similar values of the explanatory variables. Therefore, appropriate penalty matrices $P_T \in \mathbb{R}^{K_T \times K_T}$ and $P_j \in \mathbb{R}^{K_j \times K_j}$ are imposed on the basis functions defined in Equation 10. The penalty matrix for the Kronecker product of the basis functions is defined via $P_{Tj} = (\lambda_T P_j \otimes \mathbf{1}_{K_T} + \lambda_j \mathbf{1}_{K_j} \otimes P_T)$, where $\lambda_T \geq 0$ and $\lambda_j \geq 0$ denote smoothing parameters and $\mathbf{1}$ denotes the identity matrix.
As an example, we give the partial transformation function for the explanatory variable sex influencing the survival time transformation:

$$h_{\text{sex}}(t_\iota|\text{sex}) = \left(b_{\text{sex}}^{\text{lin}}(\text{sex})^\top \otimes b_T(t_\iota)^\top\right)\gamma_{\text{sex}}.$$

Since the explanatory variable sex is binary, we chose linear basis functions for $b_{\text{sex}}^{\text{lin}}(\text{sex})$, and furthermore, we chose B-spline basis functions for b_T. No penalty term P_{sex} is specified for the linear basis $b_{\text{sex}}^{\text{lin}}$ and a smoothness penalty term based on second order differences P_T is defined for the B-spline basis b_T. The resulting interaction surface for the explanatory variable sex and the survival time can also be understood as the separate estimation of a smooth survival time transformation for males and females. Hence, the difference in the survival probabilities of males and females is allowed to vary flexibly over time and is therefore able to display non-proportional hazards for the explanatory variable sex. For further details on parametrisation and penalty specification, see Hothorn et al. (2014).

Component-wise boosting algorithm for conditional transformation models for survival data The component-wise boosting algorithm for right-censored survival data is only a slight modification of the algorithm presented in Hothorn et al. (2014):

(Init) Initialise the parameters $\gamma_j^{[0]} \equiv 0$ for $j = 1, \ldots, J$, the step-size $\nu \in (0,1)$ and the smoothing parameters λ_j, $j = 1, \ldots, J$. Define the grid $t_1 < \tilde{T}_{(1)} < \ldots < \tilde{T}_{(N)} \leq t_n$. Calculate the inverse probability of censoring weights $\omega_{i\iota}$ for each grid point ι and each observation i.
Set $m = 0$.

(Gradient) Compute the negative gradient:

$$\begin{aligned} U_{i\iota} &:= -\frac{\partial}{\partial h}\rho((\tilde{T}_i \leq t_\iota, x_i), h)\Big|_{h=\hat{h}_{i\iota}^{[m]}} \\ &:= \left\{ I(\tilde{T}_i \leq t_\iota, \delta_i = 1)\frac{F'(h(t_\iota|x_i))}{F(h(t_\iota|x_i))} \cdot \frac{1}{\hat{K}(\tilde{T}_i)} \right. \\ &\quad \left. - I(\tilde{T}_i > t_\iota)\frac{F'(h(t_\iota|x_i))}{1 - F(h(t_\iota|x_i))} \cdot \frac{1}{\hat{K}(t_\iota)} \right\}\Big|_{h=\hat{h}_{i\iota}^{[m]}}, \end{aligned}$$

where $F^{'}(\cdot)$ denotes the density of the link function F, $\hat{K}(\cdot)$ denotes the marginal Kaplan-Meier estimator of the censoring distribution and

$$\hat{h}_{i\iota}^{[m]} = \sum_{j=1}^{J} \hat{h}_{j}^{[m]}(t_\iota|x_i) = \sum_{j=1}^{J} \left(b_j(x_i)^\top \otimes b_T(t_\iota)^\top \right) \gamma_j^{[m]}.$$

Fit the base-learners for $j = 1, \ldots, J$:

$$\hat{\beta}_j = \underset{\beta \in \mathbb{R}^{K_j \cdot K_T}}{arg\,min} \sum_{i=1}^{N} \sum_{\iota=1}^{n} \omega_{i\iota} \left\{ U_{i\iota} - \left(b_j(x_i)^\top \otimes b_T(t_\iota)^\top \right) \beta \right\}^2 + \beta^\top P_{Tj} \beta$$

with penalty matrix P_{Tj}.
Select the base-learner

$$j^* = \underset{j=1,\ldots,J}{arg\,min} \sum_{i=1}^{N} \sum_{\iota=1}^{n} \omega_{i\iota} \left\{ U_{i\iota} - \left(b_j(x_i)^\top \otimes b_T(t_\iota)^\top \right) \hat{\beta}_j \right\}^2.$$

(Update) the parameters $\gamma_{j^*}^{[m+1]} = \gamma_{j^*}^{[m]} + \nu \cdot \hat{\beta}_{j^*}$ and keep all other parameters fixed, i. e. $\gamma_j^{[m+1]} = \gamma_j^{[m]}$, $j \neq j^*$.

Iterate (Gradient) and (Update).

(Stop) if $m = M$. Output the model

$$\begin{aligned} \hat{\mathbb{P}}(T \leq t|\mathbf{X} = x) &= F(\hat{h}^{[M]}(t|x)) = F\left(\sum_{j=1}^{J} \hat{h}_j^{[M]}(t|x) \right) \\ &= F\left(\sum_{j=1}^{J} \left(b_j(x)^\top \otimes b_T(t)^\top \right) \gamma_j^{[M]} \right) \end{aligned}$$

as a function of arbitrary $t \in \mathbb{R}^+$ and arbitrary explanatory variables x.

3 Simulation

In the following simulations, we compared the performance of the CTM, the (stratified) Cox model, the Kaplan-Meier estimator and conditional random forests (Cforest), which estimate the conditional survivor functions nonparametrically. Thereby, we considered different scenarios of expalantory variables and proportional as well as non-proportional hazard settings. In simulation settings 1 and 2, we compared

two treatment groups $G1$ and $G2$ that differed with respect to their survival probabilities. This simulation setting has already been analysed in Schemper (1992). Moreover, we included a non-informative continuous explanatory variable x which had no influence on the survival probabilities. The two treatment groups followed the proportional hazards (PH) assumption in Simulation 1, whereas the PH assumption was violated in Simulation 2. The survival probabilities differed with respect to the treatment groups $G1$ and $G2$ *and* with respect to the continuous explanatory variable x in simulation settings 3 and 4. Again, the PH assumption was fullfilled in Simulation 3, whereas the PH assumption was violated in Simulation 4.

3.1 Simulation study setup

We investigated the performance of CTMs in comparison to alternative semi-parametric (ordinary and stratified Cox model) or non-parametric (Kaplan-Meier estimator; conditional random forests) modelling strategies in four different simulation settings with Weibull distributed survival times. Since the handling of censored observations is an important issue, we considered different amounts of right-censored survival times. The censoring times were drawn from a uniform distribution on $[0,\tau]$, $C \sim U[0,\tau]$, where the parameter τ was chosen such that 5%, 10%, 25% and 50% right-censored observations resulted in each simulation setting.

The true hazard function and the corresponding true survivor function for Weibull distributed survival times are

$$\lambda(t) = \frac{c}{b^c} t^{c-1} \quad \text{and} \quad S(t) = \exp(-b^{-c} t^c), \tag{11}$$

where b and c denote the scale and shape parameter of the Weibull distribution. The choice of parameters b and c determine whether proportional hazards or non-proportional hazards result. The PH assumption is fullfilled if the explanatory variables influence only the scale parameter b and the shape parameter c is fixed. If the explanatory variables additionally influence the shape parameter c, the PH assumption is violated, what e. g. results in crossing survivor functions.

Simulation 1 In the first simulation setting, we considered the simple data setting of two treatment groups $G1$ and $G2$ that differed with respect to their survival probabilites. The continuous explanatory variable x was uniformly distributed on $[-2,2]$ and was non-informative. The survival times were Weibull distributed with $b_1 = 1$ and $c_1 = 3$ for treatment group $G1$ and $b_2 = 1.5$ and $c_2 = 3$ for treatment group $G2$. Since the shape parameters were identical, the corresponding survivor functions followed the PH assumption (Figure 1). We sampled $N = 200$ survival times T from

the respective Weibull distribution for each treatment group and the non-informative continuous explanatory variable x was sampled independently from $x \sim U[-2,2]$.

Simulation 2 In analogy to Simulation 1, the survival probabilites differed for the treatment groups $G1$ and $G2$ and the continuous explanatory variable x was non-informative. The parameters of the Weibull distributed survival times were chosen to be $b_1 = 1.5$ and $c_1 = 3$ for treatment group $G1$ and $b_2 = 1$ and $c_2 = 1$ for treatment group $G2$. Since the scale parameters b_i and the shape parameters c_i, $i \in \{1,2\}$, were treatment-specific, the PH assumption was violated in Simulation 2 (Figure 1). We sampled $N = 200$ survival times for each treatment group and the continuous explanatory variable x was sampled independently from $x \sim U[-2,2]$.

Simulation 3 In contrast to the previous simulation settings, the survival times differed with respect to the treatment group *and* with respect to the continuous explanatory variable x in this setting. The survival times were Weibull distributed with scale parameters $b_1 = \exp(\frac{1}{4} + x)$ for treatment group $G1$ and $b_2 = \exp(1 + x)$ for treatment group $G2$, where x was uniformly distributed on $[0,1]$. Identical shape parameters were chosen, $c_1 = c_2 = 3$, what resulted in the PH assumption. The connection to the Cox model can be established by rewriting the conditional Weibull distribution in terms of the Cox linear transformation model (Equation 3). The conditional Weibull distribution results from Equation 11 by inserting the scale parameter $b = \exp(\beta_G + x)$, where $\beta_G = \frac{1}{4}$ for $G1$ and $\beta_G = 1$ for $G2$, and the shape parameter $c = 3$:

$$\begin{aligned} 1 - S(t|G,x) &= 1 - \exp(-\exp(\beta_G + x)^{-3} \cdot t^3) \\ &= 1 - \exp(-\exp(-3 \cdot (\beta_G + x) + 3 \cdot \log(t))) \\ &= 1 - \exp(-\exp(\tilde{\beta}_G + \tilde{\beta}_x \cdot x + h_T(t))) \\ &= F(h_T(t) + \tilde{\mathbf{x}}^\top \tilde{\beta}), \end{aligned} \tag{12}$$

where F denotes the minimum-extreme value distribution, $\tilde{\beta} = (\beta_G \ \beta_x)^\top$ and $\tilde{\mathbf{x}} = (G \ x)$. More precisely, in this simulation setting the parameters of the linear transformation model were $\tilde{\beta}_G = -\frac{3}{4}$ for $G1$ and $\tilde{\beta}_G = -3$ for $G2$, $\tilde{\beta}_x = -3$ and $h_T(t) = 3 \cdot \log(t)$. Hence, the simulation setting could be perfectly analysed using a Cox model, since there were no interaction terms of the explanatory variables and the survival time and G and x had a linear influence. The simulation study was based on $N = 600$ observations. Thereby, we first sampled 600 observations for $x \sim U[0,1]$. Afterwards, we sampled 300 observations from the Weibull distribution with parameters b_1 and c and 300 observations from the Weibull distribution with parameters b_2 and c using one half of the x-values, respectively.

Figure 1: Simulation: True survivor and hazard functions for treatment groups $G1$ and $G2$ based on Weibull distributed survival times. Proportional hazards setting: $b_1 = 1$, $c_1 = 3$ (for $G1$) and $b_2 = 1.5$, $c_2 = 3$ (for $G2$); Non-proportional hazards setting: $b_1 = 1.5$, $c_1 = 3$ and $b_2 = 1$, $c_2 = 1$.

Simulation 4 In anology to Simulation 3, the survival probabilities were influenced by *G and x*. But this time, we chose a non-proportional hazards setting by keeping the scale parameter $b = \exp(\frac{1}{2})$ fixed and letting the shape parameter depend on the explanatory variables: $c_1 = 2 + x^2$ for treatment group $G1$ and $c_2 = 2.5 + x^2$ for treatment group $G2$. Hence, the shape parameters differed only slightly for the treatment groups and were mainly influenced non-linearly by x.

Again, the corresponding conditional Weibull distribution of the survival times can be displayed as a conditional transformation model:

$$1 - S(t|G,x) = 1 - \exp\left(-\exp\left(-\frac{1}{2}\cdot\beta_G - \frac{1}{2}x^2 + \beta_G\cdot\log(t) + x^2\cdot\log(t)\right)\right),$$

where $\beta_G = 2$ for treatment group $G1$ and $\beta_G = 2.5$ for treatment group $G2$. Since there were interaction terms between the explanatory variables and the survival time, the survivor functions did not fullfill the PH assumption. We first sampled $N = 600$ observations for $x \sim U[0,2]$. Afterwards, 300 observations were sampled from the Weibull distribution with parameters b and c_1 and 300 observations were sampled from the Weibull distribution with parameters b and c_2 using one half of the x-values, respectively.

3.2 Model estimation

Simulation 1 We estimated the conditional survival curves for the treatment groups $G1$ and $G2$ and the continuous covariate x, $S(t|G,x)$, using a CTM, an ordinary Cox model, the Kaplan-Meier estimator, and conditional random forests. Thereby, x could be considered in the CTM, the Cox model and in conditional random forests. The conditional Kaplan-Meier estimator can only be obtained for categorical explanatory variables and separate estimates were obtained for the treatment groups. Therefore, the conditional Kaplan-Meier estimator can be understood as a nonparametric alternative to conditional random forests, whereby the Kaplan-Meier estimator is supposed to perform better, since the non-informative explanatory variable x is ignored. In the Cox model, the hazard function is modelled via $\lambda(t|G,x) = \lambda_0(t)\exp(\beta_G\cdot G + \beta_x\cdot x)$, whereby proportional hazards are assumed and $\lambda_0(t)$ is the baseline hazard. In the CTM, the conditional transformation function consists of a partial transformation function for each explanatory variable, $h(t|G,x) = h_G(t|G) + h_x(t|x)$, whereby the influence of the explanatory variables may vary over time.

Simulation 2 Since simulation setting 2 is a non-proportional hazards setting, the conditional survival curves $S(t|G,x)$ were additionally estimated using a stratified Cox model. The continuous covariate x could be considered in the CTM, in the (ordinary and stratified) Cox model and in conditional random forests. Since x is non-informative in this setting, treatment-specific Kaplan-Meier estimators were obtained as a non-parametric and predominant alternative to conditional random forests. The hazard function was modelled via $\lambda(t|G,x) = \lambda_0(t)\exp(\beta_G\cdot G + \beta_x\cdot x)$ in the Cox model, whereas treatment-specific hazard functions were assumed in the

stratified Cox model, $\lambda(t|G,x) = \lambda_G(t)\exp(\beta_x \cdot x)$. The influence of the explanatory variables might vary over time in the CTM: $h(t|G,x) = h_G(t|G) + h_x(t|x)$.

Simulation 3 Simulation setting 3 assumed proportional hazards and linear influences for both explanatory variables G and x on the survival time, which perfectly fits a Cox model. Therefore, we estimated the conditional survivor functions $S(t|G,x)$ using a CTM, a Cox model and conditional random forests. The hazard function is modelled via $\lambda(t|G,x) = \lambda_0(t)\exp(\beta_G \cdot G + \beta_x \cdot x)$ in the Cox model. Since the conditional survivor function can also be written in terms of a linear transformation model (Equation 12), we restricted the flexibility of the CTM to the flexibility of a Cox model: $h(t|G,x) = h_G(1|G) + h_x(1|x) + h_T(t|1) = \beta_G \cdot G + \beta_x \cdot x + h_T(t)$. Hence, we avoided any interactions between the explanatory variables and the survival time and we assumed linear influences for G and x.

Simulation 4 In Simulation 4, we assumed non-proportional hazards and G and x were influential. The conditional survivor functions $S(t|G,x)$ were estimated using a CTM, an ordinary and a stratified Cox model and conditional random forests. The hazard function was estiamted via $\lambda(t|G,x) = \lambda_0(t)\exp(\beta_G \cdot G + \beta_x \cdot x)$ in the ordinary Cox model. In the stratified Cox model, we assumed treatment-specific baseline hazards: $\lambda(t|G,x) = \lambda_G(t)\exp(\beta_x \cdot x)$. Nevertheless, the stratified Cox model still ignores the interaction between the survival time and x. In the CTM, we chose the flexible conditional transformation function $h(t|G,x) = h_G(t|G) + h_x(t|x)$.

3.3 Model evaluation

We aimed at evaluating the goodness of the CTM, the (ordinary and stratified) Cox model, the Kaplan-Meier estimator and conditional random forests for estimating the survivor functions of treatment groups $G1$ and $G2$ in all four simulation settings. Therefore, we used the out-of-sample uncensored log score (Equation 7) and the mean absolute deviation (MAD) between the true and the estimated survivor functions as quality criteria.

For the evaluation in simulation settings 1 and 2, we drew 100 new observations from the (known) Weibull distributions for each treatment group and calculated separate uncensored log scores for the two treatment groups afterwards. For example, we describe the calculation of the uncensored log score for treatment group G_1: We compared the true survivor status $I(T_l \leq t_\iota)$ for each new survival time $T_l, l = 1,\ldots,100$, for treatment group G_1 along a grid of time points t_ι with the corresponding estimated survival probabilities $\pi(t_\iota|G_1,x_i)$ for the $i = 1,\ldots,200$

observations in treatment group G_1. Thereby, the conditional survival probabilities $\pi(t_\iota|G_1,x_i)$ were the estimated survival probabilities resulting from the CTM (where $\pi(t_\iota|G_1,x_i) = F(\hat{h}_G(t_\iota|G_1) + \hat{h}_x(t_\iota|x_i))$), the (ordinary or stratified) Cox model, or conditional random forests. The survival probabilites $\pi(t_\iota|G_1)$ were only treatment-specific for the Kaplan-Meier estimator. The uncensored log score for treatment group G_2 was calculated analogously.

The calculation of the uncensored log score had to be adapted for simulation settings 3 and 4. Since the continuous explanatory variable x was influential, we drew one new observation from the corresponding Weibull distribution for each G–x combination in the simulation data set. Hence, the evaluation was based on 600 new observations, 300 new observations for each treatment group. The uncensored log score was calculated treatment-specific again. For example, we drew a new observation T_{new} from the Weibull distribution with distribution parameters depending on treatment group $G1$ and $x = 0.5$. The binary survivor status $I(T_{\text{new}} \leq t_\iota)$ along a grid of time points t_ι was compared with the corresponding estimated survival probabilities $\pi(t_\iota|G_1, x = 0.5)$.

In addition, we calculated the MAD of the estimated survival curves and the true Weibull distribution functions for each treatment group separately:

$$\text{MAD}(G_k) = \frac{1}{n \cdot \frac{N}{2}} \sum_{\iota=1}^{n} \sum_{i=1}^{\frac{N}{2}} |p(t_\iota|G_k,x_i) - \pi(t_\iota|G_k,x_i)|, \tag{13}$$

where p denotes the true survival probabilities and π denotes the estimated survival probabilities. Furthermore, $k \in 1,2$ denotes the index for the two treatment groups and $i = 1,\ldots,\frac{N}{2}$ is the index for the observations in each treatment group. In the simulation settings 1 and 2, the true survival probabilities $p(t_\iota|G_k,x_i)$ reduced to $p(t_\iota|G_k)$, since x was non-informative. For reasons of interpretability, the MAD values and the uncensored log scores were multiplied by 100. We evaluated the MADs and the uncensored log scores on a grid of time points consisting of all censoring and event time points.

This procedure was repeated for $B = 100$ simulated data sets. We calculated mean values of the resulting 100 MADs or uncensored log scores for the different treatment groups and the different estimation techniques.

Simulation 1 In the proportional hazards setting with non-informative explanatory variable x, all four estimation approaches yielded similar results. The calculated mean MAD values (Table 1; Figure 2) were small for all model approaches and indicated, that the estimated survivor functions were in good accordance with the true Weibull survivor functions for all amounts of censoring. The Cox model and the Kaplan-Meier estimator performed slightly better, since the Cox model

profited from the PH assumption and the Kaplan-Meier estimator ignored the non-informative covariate x. Nevertheless, the uncensored log score was the more interesting quality criterion as it evaluates how well the estimation techniques are able to predict the survivor status of *new* observations. Again, all four estimation approaches yielded similar results (Table 2; Figure 3).

Table 1: Simulation 1: Mean absolute deviations between true and estimated survival curves for each treatment group. The reported values are mean values over $B = 100$ simulations.

Treatment group	Model	Censoring			
		5%	10%	25%	50%
G1	CTM	2.32	2.40	2.54	2.91
	Cox	1.95	2.04	2.11	2.40
	Kaplan-Meier	1.86	1.93	2.04	2.31
	Cforest	2.02	2.09	2.14	2.50
G2	CTM	2.36	2.39	2.36	2.42
	Cox	1.90	1.91	1.88	1.94
	Kaplan-Meier	2.03	2.04	2.04	2.05
	Cforest	2.23	2.25	2.21	2.24

Table 2: Simulation 1: Out-of-sample uncensored log score based on 100 new observations for each treatment group. The reported values are mean values over $B = 100$ simulations.

Treatment group	Model	Censoring			
		5%	10%	25%	50%
G1	CTM	42.97	42.82	42.15	41.02
	Cox	42.74	42.59	41.85	40.45
	Kaplan-Meier	43.27	43.15	42.60	42.17
	Cforest	43.32	43.19	42.61	41.64
G2	CTM	50.35	49.32	45.93	40.66
	Cox	50.18	49.12	45.66	40.19
	Kaplan-Meier	50.33	49.29	45.91	40.59
	Cforest	50.42	49.39	45.99	40.68

Simulation 2 In the non-proportional hazards setting with non-informative covariate x, the MADs of the CTM, the stratified Cox model, the Kaplan-Meier estimator and conditional random forests were similar throughout, whereas the ordi-

Figure 2: Simulation 1: Boxplot of the treatment-specific mean MAD values based on $B = 100$ simulations for the conditional transformation model (CTM), the Cox model (Cox), the Kaplan-Meier estimator (KM), and conditional random forests (Cforest). 5%, 10%, 25%, and 50% of right-censored observations were observed.

nary Cox model clearly yielded higher MADs (Table 3; Figure 4). The only exception were the MAD values of treatment group $G2$ for 50% censored observations, where all models, but especially the CTM model, had higher MADs. Moreover, the MAD values for conditional random forests were most variable. The calculated uncensored log scores gave similar results (Table 4; Figure 5). Again, the log scores for the CTM, the stratified Cox model, the Kaplan-Meier estimator and conditional random forests were almost equal, whereas the ordinary Cox model clearly yielded higher values. One exception was the Kaplan-Meier estimator for treatment group $G2$ for 50% of censored observations, which yielded worse results. This might be due to the Kaplan-Meier estimator resulting in a step function with abrupt steps for risk sets containing few observations.

Simulation 3 The Cox model and the CTM approach performed comparably well in the proportional hazards setting with influential explanatory variables x and G. The mean MAD values (Table 5; Figure 6) and the out-of-sample uncensored log

Figure 3: Simulation 1: Boxplot of the out-of-sample mean uncensored log scores based on 100 new observations for each treatment group and $B = 100$ simulations for the conditional transformation model (CTM), the Cox model (Cox), the Kaplan-Meier estimator (KM), and conditional random forests (Cforest). 5%, 10%, 25%, and 50% of right-censored observations were observed.

scores (Table 6; Figure 7) were similar for the Cox model and the CTM, whereas conditional random forest predictions were associated with higher MAD values and uncensored log scores. This could be due to the fact, that conditional random forests were not able to profit from the PH assumption in this setting.

Simulation 4 In the non-proportional hazards setting with influential explanatory variables G and x, the CTM performed better than all alternative modelling approaches. Hence, the CTM approach showed lower MAD values for all amounts of censoring than the Cox model, the stratified Cox model and conditional random forests (Table 7; Figure 8). The differences were minor for the out-of-sample mean uncensored log scores, but nevertheless, the CTM approach was associated with the smallest mean uncensored log scores (Table 8; Figure 9). This is due to the fact, that the CTM approach was the only approach that was able to account for the non-linear influence of x on the shape parameter of the Weibull distribution

Table 3: Simulation 2: Mean absolute deviations between true and estimated survival curves for each treatment group. The reported values are mean values over $B = 100$ simulations.

Treatment group	Model	Censoring			
		5%	10%	25%	50%
G1	CTM	1.95	1.97	2.02	2.26
	Cox	9.38	9.30	8.99	5.65
	Kaplan-Meier	1.71	1.73	1.76	1.76
	Cforest	3.19	2.92	2.36	1.88
	Stratified Cox	1.92	1.95	1.96	1.96
G2	CTM	2.99	3.07	3.34	5.83
	Cox	8.56	8.53	8.24	6.68
	Kaplan-Meier	2.30	2.36	2.49	3.61
	Cforest	3.92	3.78	3.41	3.38
	Stratified Cox	2.63	2.71	2.86	3.49

Table 4: Simulation 2: Out-of-sample uncensored log score based on 100 new observations for each treatment group. The reported values are mean values over $B = 100$ simulations.

Treatment group	Model	Censoring			
		5%	10%	25%	50%
G1	CTM	38.63	38.16	36.67	32.47
	Cox	42.10	41.62	40.05	34.20
	Kaplan-Meier	38.75	38.27	36.79	32.44
	Cforest	38.98	38.42	36.76	32.45
	Stratified Cox	38.65	38.19	36.69	32.41
G2	CTM	51.10	51.66	53.31	59.16
	Cox	55.87	56.46	57.85	60.26
	Kaplan-Meier	51.02	51.67	53.37	64.46
	Cforest	52.26	52.73	53.97	57.95
	Stratified Cox	51.13	51.78	53.49	57.74

adequately. The Cox model performed worst, since it was not able to consider non-proportional hazards. The stratified Cox model considered only non-proportional hazards in G, but was not able to consider non-proportional hazards in x. Since the non-proportionality of hazards was mainly induced by x, the stratified Cox model performed only slightly better than the ordinary Cox model. Conditional random forests were able to account for non-proportional hazards in G and x by searching for adequate split points. Therefore, conditional random forests performed better

Figure 4: Simulation 2: Boxplot of the treatment-specific mean MAD values based on $B = 100$ simulations for the conditional transformation model (CTM), the Cox model (Cox), the Kaplan-Meier estimator (KM), conditional random forests (Cforest), and the stratified Cox model (Cox.Strata). 5%, 10%, 25%, and 50% of right-censored observations were observed.

Table 5: Simulation 3: Mean absolute deviations between true and estimated survival curves for each treatment group. The reported values are mean values over $B = 100$ simulations.

Treatment group	Model	Censoring			
		5%	10%	25%	50%
G1	CTM	1.34	1.37	1.54	2.01
	Cox	1.21	1.25	1.36	1.58
	Cforest	2.97	3.11	3.47	4.13
G2	CTM	1.27	1.26	1.27	1.36
	Cox	1.25	1.25	1.25	1.28
	Cforest	3.22	3.20	3.01	2.92

than the ordinary and the stratified Cox model, but worse than the CTM, since the influence of x on the shape parameter varied non-linearly.

Figure 5: Simulation 2: Boxplot of the out-of-sample mean uncensored log scores based on 100 new observations for each treatment group and $B = 100$ simulations for the conditional transformation model (CTM), the Cox model (Cox), the Kaplan-Meier estimator (KM), conditional random forests (Cforest), and the stratified Cox model (Cox.Strata). 5%, 10%, 25%, and 50% of right-censored observations were observed.

4 Chronic myelogenous leukaemia data

Curative bone marrow transplantation is feasible for only a minority of patients with chronic myelogenous leukaemia. Therefore, drug-based chemotherapy remains a treatment of central interest. The standard chemotherapy has long been with the cytostatic drugs busulfan (BUS) or hydroxyurea (HU). In a multicentre, randomised study, Hehlmann, Heimpel, Hasford, and Others (1994) have shown that treatment with the drug interferon-α (IFN-α) significantly prolongs survival compared to treatment with BUS, and survival times after treatment with IFN-α or HU were not significantly different. Within the scope of the study, 516 eligible patients were recruited in 57 study centres from 1983 to 1991. For 507 of the 516 patients, complete data on sex, age and a prognostic score distinguishing between low, intermediate and high risk groups (Hasford, Pfirrmann, Hehlmann, and Oth-

Table 6: Simulation 3: Out-of-sample uncensored log score based on one new observation for each combination of the continuous explanatory variable x and treatment group G (results in 600 new observations). The reported values are mean values over $B = 100$ simulations.

Treatment group	Model	Censoring			
		5%	10%	25%	50%
G1	CTM	41.49	41.38	40.57	41.20
	Cox	41.43	41.30	39.60	38.84
	Cforest	43.69	43.31	42.38	41.90
G2	CTM	31.51	30.83	28.05	24.90
	Cox	31.51	30.83	28.07	24.93
	Cforest	32.52	31.82	29.08	26.00

Figure 6: Simulation 3: Boxplot of the treatment-specific mean MAD values based on $B = 100$ simulations for the conditional transformation model (CTM), the Cox model (Cox), and conditional random forests (Cforest). 5%, 10%, 25%, and 50% of right-censored observations were observed.

ers, 1998) are available. Of the 507 patients, 132 random patients were treated with

Figure 7: Simulation 3: Boxplot of the out-of-sample mean uncensored log scores based on 300 new observations for each treatment group and $B = 100$ simulations for the conditional transformation model (CTM), the Cox model (Cox), and conditional random forests (Cforest). 5%, 10%, 25%, and 50% of right-censored observations were observed.

IFN-α, 182 were treated with BUS and 193 were treated with HU. 90 patients were right-censored mainly due to bone marrow transplantation during the first chronic phase, and 417 patients died during the study period (Herberich and Hothorn, 2012).

Herberich and Hothorn (2012) analysed the treatment effects using a frailty Cox model (McGilchrist and Aisbett, 1991b) with Gaussian frailties for the 57 study centres. Furthermore, age, sex, treatment and risk group were included as linear predictors. We slightly modified the frailty Cox model by including the interaction between treatment and risk group instead of including both predictors additively. This resulted in nine treatment–risk group combinations, whereby one regression coefficient is estimated for each combination in the frailty Cox model. Hence, the conditional hazard function in the frailty Cox model given the explanatory variables x is

$$\lambda(t_\iota|\mathbf{x}) = Z \cdot \lambda_0(t_\iota) \cdot \exp(\beta_{\text{tr:risk}} x_{\text{tr:risk}} + \beta_{\text{age}} x_{\text{age}} + \beta_{\text{sex}} x_{\text{sex}}),$$

Table 7: Simulation 4: Mean absolute deviations between true and estimated survival curves for each treatment group. The reported values are mean values over $B = 100$ simulations.

Treatment group	Model	Censoring			
		5%	10%	25%	50%
G1	CTM	3.17	3.16	3.15	3.15
	Cox	5.67	5.60	5.34	4.78
	Cforest	4.35	4.42	4.53	4.40
	Stratified Cox	5.57	5.51	5.28	4.75
G2	CTM	2.91	2.93	2.91	2.84
	Cox	4.93	4.87	4.60	4.07
	Cforest	3.69	3.76	3.80	3.65
	Stratified Cox	4.93	4.87	4.61	4.09

Table 8: Simulation 4: Out-of-sample uncensored log score based on one new observation for each combination of the continuous explanatory variable x and treatment group G (results in 600 new observations). The reported values are mean values over $B = 100$ simulations..

Treatment group	Model	Censoring			
		5%	10%	25%	50%
G1	CTM	48.84	47.77	44.97	40.62
	Cox	49.91	48.86	45.98	41.37
	Cforest	49.24	48.26	45.54	41.15
	Stratified Cox	49.94	48.90	46.04	41.46
G2	CTM	45.88	44.83	41.82	37.13
	Cox	46.91	45.79	42.70	37.83
	Cforest	46.38	45.31	42.33	37.59
	Stratified Cox	46.76	45.65	42.58	37.77

where $\lambda_0(\cdot)$ denotes the baseline hazard function and Z denotes the frailty term. Since the frailty Cox model assumes proportional hazards for all patient characteristics, we alternatively used a CTM for data analysis. In the CTM, we allowed for flexible influences of each treatment–risk group combination, sex and age over time and thereby the proportional hazards assumption is relaxed. We defined separate partial transformation functions for sex, age, the treatment-risk group interaction and the study centers, what resulted in the conditional transformation function:

$$h(t_\iota|\mathbf{x}) = h_{\text{tr:risk}}(t_\iota|\text{tr:risk}) + h_{\text{sex}}(t_\iota|\text{sex}) + h_{\text{age}}(t_\iota|\text{age}) + h_{\text{centre}}(1|\text{centre}),$$

Figure 8: Simulation 4: Boxplot of the treatment-specific mean MAD values based on $B = 100$ simulations for the conditional transformation model (CTM), the Cox model (Cox), conditional random forests (Cforest), and the stratified Cox model (Cox.Strata). 5%, 10%, 25%, and 50% of right-censored observations were observed.

whereby the partial transformation functions were specified using basis functions (Equation 10): $h_{\text{tr:risk}}(t_\iota|\text{tr:risk}) = \left(b^{\text{lin}}_{\text{tr:risk}}(\text{tr:risk})^\top \otimes b_T(t_\iota)^\top\right)\gamma_{\text{tr:risk}}$, $h_{\text{sex}}(t_\iota|\text{sex}) = \left(b^{\text{lin}}_{\text{sex}}(\text{sex})^\top \otimes b_T(t_\iota)^\top\right)\gamma_{\text{sex}}$, $h_{\text{age}}(t_\iota|\text{age}) = \left(b_{\text{age}}(\text{age})^\top \otimes b_T(t_\iota)^\top\right)\gamma_{\text{age}}$ and $h_{\text{centre}}(1|\text{centre}) = b^{\text{lin}}_{\text{centre}}(\text{centre}) \otimes b^{\text{lin}}(\mathbf{1})$. In other words, we fitted a separate function over time for each treatment–risk group combination and for both sexes. For the age effect, we estimated a bivariate interaction surface depending on age and survival times. Moreover, we included a penalised parameter for the different study centres, that was assumed to be constant over time. By including the treatment–risk group interaction, we investigated whether the superiority of IFN-α therapy occurs in all three risk groups, or whether different treatments should be considered depending on the specific risk group.

Model estimation. We analysed the chronic myelogenous leukaemia data using both a frailty Cox model and a CTM with a more flexible modelling approach. In

Figure 9: Simulation 4: Boxplot of the out-of-sample mean uncensored log scores based on 300 new observations for each treatment group and $B = 100$ simulations for the conditional transformation model (CTM), the Cox model (Cox), conditional random forests (Cforest), and the stratified Cox model (Cox.Strata). 5%, 10%, 25%, and 50% of right-censored observations were observed.

the frailty Cox model, the patient characteristics sex, age, treatment, and risk group and the interaction treatment–risk group were included as linear explanatory variable effects. The estimated risks (resulting from the exponential transformation of the estimated coefficients) for the explanatory variables (Table 9) can be interpreted as multiplicative effects on the hazard rate compared to the reference category. The shape of the corresponding estimated survival curves (Figure 10) was partly determined by the proportional hazards assumption. The treatment IFN-α in the low risk group served as reference category for the treatment–risk group interaction. In the low risk group, patients treated with IFN-α had the lowest hazard rate and thus the highest survival probabilities, followed by patients treated with BUS, followed by patients treated with HU. In the intermediate risk group, patients treated with IFN-α had the highest survival probabilities, followed by patients treated with HU; patients treated with BUS clearly had the worst hazard rates. In the high risk group, again patients treated with IFN-α had the lowest hazard rate; patients treated with

HU or BUS had higher and almost equal hazard rates. Within all risk groups, the differences between the three treatments are not significant on the 5% significance level, but the differences between the risk groups are significant. In summary, the inclusion of the interaction term was important as it revealed that BUS treatment might be superior to HU treatment in the low risk group, whereas HU treatment might be superior to BUS treatment in the intermediate risk group. IFN-α treatment seemed to be superior in all risk groups, but this effect is not significant.
Moreover, females had a 0.8-times lower hazard rate than males and therefore higher survival times. Age had no influence on the survival probabilities.

Table 9: Estimated risks (exp(coef)) and corresponding standard deviations (sd) resulting from the frailty Cox model.

Patient characteristic	exp(coef) (sd)
Age	1.00 (0.004)
Sex	0.79 (0.102)
Treatment BUS : Risk group 0	1.39 (0.253)
Treatment HU : Risk group 0	1.58 (0.245)
Treatment IFN : Risk group 1	1.68 (0.251)
Treatment BUS : Risk group 1	2.57 (0.314)
Treatment HU : Risk group 1	1.87 (0.306)
Treatment IFN : Risk group 2	2.06 (0.295)
Treatment BUS : Risk group 2	3.48 (0.371)
Treatment HU : Risk group 2	3.47 (0.368)

When we estimated CTMs, we obtained a survival curve for each patient in the data set, and the curve was dependent on the individual's sex, age, treatment–risk group combination and study centre. The estimated survival curves resulting from a CTM delivered separate survival curves for each category (Figure 11). The remaining predictors were set to their reference category (sex = male, risk group = low, treatment = IFN-α, study centre = 1) and age was set to 45. When we calculated survival curves for age, we chose ages 36, 48 and 58, which represent the 25%-quantile, the median, and the 75%-quantile, respectively, for the patients in the data set.
When we considered the treatment–risk group combinations, the three survival curves belonging to the low risk group (solid lines) showed the highest survival probabilities. In this case, treatment with IFN-α was superior to treatment with HU or with BUS at early time points, and the survival curves for HU treatment and BUS treatment were almost identical. In contrast, at late time points, the survival probabilities for BUS treatment are higher than those for IFN-α or HU treatment.

Figure 10: Estimated survival curves resulting from the frailty Cox model. Treatments with the cytostatic drugs busulfan, hydroxyurea, and interferon-α is abbreviated by BUS, HU, IFN. The three age categories represent the 25%-, 50%-, and 75%-quantile for the patients in the data set; the age categories are overplotted.

The crossing of the survival curves cannot occur in the Cox model. For the intermediate risk group, the survival curve for treatment with IFN-α is similar to that for treatment with HU, and patients treated with BUS had clearly lower survival probabilities. Patients in the high risk group treated with BUS or HU had low and almost identical survival probabilities. Patients treated with IFN-α had clearly the highest survival probabilities of patients in the high risk group and even higher survival probabilities than some of the patients in the intermediate risk group.
The estimated survival probabilities of males and females were similar at very early time points. Thereafter, females clearly had higher survival probabilities than males, in accordance with the frailty Cox model.
Considering the estimated survival curves with regard to age, patients of age 36 had slightly lower survival probabilities than patients of age 48, and patients of age 58 had the highest survival probabilities. This tendency for longer survival with increasing age was not observed in the frailty Cox model.

Model evaluation. To evaluate the performance of the CTM and the frailty Cox model for the chronic myelogenous leukaemia data set, we calculated the censored log score given in Equation 9 for a set of new observations. Hence, our aim was to evaluate the accordance of the true survival status of a set of new patients and the

Figure 11: Estimated survival curves resulting from the CTM. Treatments with the cytostatic drugs busulfan, hydroxyurea, and interferon-α is abbreviated by BUS, HU, IFN. The three age categories represent the 25%-, 50%-, and 75%-quantile for the patients in the data set.

predicted survival probabilities resulting from the proposed CTM and frailty Cox model. As there are no new patients, we used a bootstrap approach to generate a learning data set for model estimation and an evaluation data set consisting of "new" observations for model evaluation. The learning data set was generated by randomly choosing $N = 507$ patients from the original data set *with* replacement. The design of the selection procedure guaranteed that at least one patient of each study centre is included in the learning data set. This procedure generates a valid bootstrap data set as Efron (1981) has shown that if censoring occurs randomly, one can simply draw an independent sample of observed survival times *and* respective censoring indicators with replacement from the original data set. Since we sampled the observations with replacement, some of the original observations might occur manifold in the learning data set and some are not selected at all. The observations that were not selected for the learning data set were considered as "new" observations and formed the evaluation data set. To determine the censored log score for the evaluation data set, we used the following procedure:

1. Based on the learning data set, the proposed CTM and a frailty Cox model were estimated.

2. For all observations belonging to the evaluation data set, we predicted the survival probabilities $\pi(t_l|x_i)$ using the estimated CTM and the estimated frailty Cox model (Step 1). Thereby, t_l denotes an arbitrary time point from the grid consisting of all event and censoring time points of the evaluation data set.
3. The censored log score was calculated based on the predicted survival probabilities (Step 2) and the corresponding true survival status of all patients in the evaluation data set. Separate censored log scores for the CTM and the frailty Cox model resulted.

The whole procedure was repeated 50 times, resulting in 50 bootstrap replications.

The CTM approach resulted in higher censored log scores than the frailty Cox model (Figure 12), which indicated a better accordance of predicted survival probabilities and true survival status for the CTM. In agreement with this result, the Kolmogorov-Smirnov test indicated a significantly lower empirical distribution function of the CTM log scores compared to the log scores resulting from the frailty Cox model ($p < 0.01$).

5 Discussion

The direct estimation of the survivor function in survival data analysis is of special interest as the reliable prediction of patient-specific survivor functions allows a better prognosis of the course of disease (Mackillop and Quirt, 1997). We propose the use of conditional transformation models (CTMs) to directly estimate the conditional survivor function of the survival times given a set of patient characteristics.

The well-known Cox model is the regression model most commonly used in survival analysis (Cox, 1972). One important restriction of the Cox model is the proportional hazards assumption. Of course, several strategies deal with or identify non-proportional hazards for some of the explanatory variables. For example, if non-proportional hazards for a categorical variable are identified, the estimation of a stratified Cox model with separate baseline hazard functions for the subgroups is frequently used. Speculation about the validity of the proportional hazards assumption in the Cox model becomes superfluous when the CTM approach is used, because the proportional hazards assumption is relaxed and can be checked easily by graphical comparisons.

In our simulation, we investigated the performance of the CTM in cases of proportional hazards and non-proportional hazards and compared the performance to that of the (ordinary or stratified) Cox model, the Kaplan-Meier estimator and conditional random forests. We measured the performance in terms of the correspondence of true and estimated survival probabilities for new observations. In the

Figure 12: Out-of-sample censored log scores for the CTM and the frailty Cox model based on 50 bootstrap evaluation data sets.

simulation settings with informative binary treatment group and non-informative continuous explanatory variable, the CTM was able to keep up with the alternative methods in the case of proportional hazards. In the case of non-proportional hazards, the CTM clearly outperformed the ordinary Cox model and delivered results equally as good as those of the stratified Cox model, the Kaplan-Meier estimator and conditional random forests. In the simulation settings with informative binary treatment group and informative continuous explanatory variable, the CTM performed as good as the ordinary Cox model in the proportional hazards setting. In

the non-proportional hazards setting, the CTM outperformed all alternative models, since it was the only method that was able to consider non-proportionality induced non-linearly by a continuous explanatory variable. One further advantage of the CTM was that owing to the imposed smoothness penalty, smooth estimated survival curves resulted, which is more realistic than the step functions resulting from the Cox model, the Kaplan-Meier estimator and conditional random forests. Moreover, the results of the simulation study showed that the CTM can handle up to 50% of right-censored observations without heavy losses in the quality of the resulting estimates.

Furthermore, we used the CTM approach to analyse survival times of patients suffering from chronic myelogenous leukaemia and compared the results to results of the frailty Cox model, which has been used in the past. The CTM results revealed that the proportional hazards assumption is not valid for all treatment–risk group combinations. Hence, the influence of the risk group and treatment on the survival probabilities can be investigated more differentiatedly and more thoroughly with the CTM than with the frailty Cox model. The censored log score was used to assess the accurateness of the prediction of the survival probabilities for new patients and revealed a superiority of the CTM in prediction accuracy compared to the frailty Cox model.

The handling of right-censored observations is a main topic in survival analysis. In CTMs, the IPCW approach has been used to account for right-censored observations. The integrated Brier score or log score for right-censored observations are well-established scoring rules for model assessment and comparison, but, to the best of our knowledge, they have not been yet used as risk functions for model estimation. In the IPCW approach, the observations are reweighted by the inverse probability of remaining uncensored up to a specific time point. In CTMs, this probability is calculated in terms of the marginal Kaplan-Meier estimator of the censoring distribution. Hence, the weights are calculated based on observed data and, more importantly, it is assumed that the censoring mechanism does not depend on any explanatory variables. Especially the dependency of the censoring distribution on (some of) the explanatory variables would be a worthwhile extension and needs further investigation (Gerds and Schumacher, 2006). Nevertheless, Hothorn et al. (2014) showed the consistency of the conditional transformation function h in CTMs which transfers to CTMs for survival data, since we only adapted the weighting scheme to account for right censoring. Mackenzie (2012) previously estimated survival curves with dependent left truncated data using Cox's model and inverse probability weighting. Thus, it would be interesting if and how the suggested approach extends to left-truncated or interval-censored data.

Basically, three main assumptions are made when estimating CTMs for survival data. First of all, by assuming that the transformation function h exists, we

assume that there is a monotone transformation from the unknown survival time distribution to the link function F. Furthermore, h is decomposed additively into partial transformation functions, whereby additivity on the scale of the transformation function is assumed. And last, the event times and the right-censoring times are assumed to be independent, what is a strong but common assumption in survival data analysis. The user should be aware of these model assumptions, since they might be violated.

6 Software

All analyses were carried out in the R system of statistical computing (R Core Team, 2013). CTMs were estimated using the R add-on package **ctmDevel** (Hothorn, 2013). To compare the proposed CTMs for survival data with established models, we estimated Cox models using the R add-on package **survival** (Therneau, 2013) and calculated Kaplan-Meier estimators using the R add-on package **prodlim** (Gerds, 2013). R code for reproducing the results of Section 3 (in `ctmDevel/inst/empeval`) and Section 4 (in `ctmDevel/inst/applications`) is publicly available in the **ctm**-package from the R-forge repository (`https://r-forge.r-project.org/projects/ctm`).

Acknowledgments

We are grateful to the German CML Study Group (Head: Prof. Dr. R. Hehlmann) and especially to Markus Pfirrmann for providing us with the chronic myelogenous leukaemia data.

References

Aalen, O. (1988): "Heterogeneity in survival analysis," *Statistics in Medicine*, 7, 1121–1137.

Andersen, P. K., E. Christensen, L. Fauerholdt, and P. Schlichting (1983): "Measuring prognosis using the proportional hazards model," *Scandinavian Journal of Statistics*, 10, 49–52.

Bühlmann, P. and T. Hothorn (2007): "Boosting algorithms: Regularization, prediction and model fitting," *Statistical Science*, 22, 477–505.

Chen, K., Z. Jin, and Z. Ying (2002): "Semiparametric analysis of transformation models with censored data," *Biometrika*, 89, 659–668.

Cheng, S. C., L. J. Wei, and Z. Ying (1995): "Analysis of transformation models with censored data," *Biometrika*, 82, 835–845.

Cheng, S. C., L. J. Wei, and Z. Ying (1997): "Predicting survival probabilities with semiparametric transformation models," *Journal of the American Statistical Association*, 92, 227–235.

Chernozhukov, V. and H. Hong (2002): "Three-step censored quantile regression and extramarital affairs," *Journal of the American Statistical Association*, 97, 872–882.

Cox, D. R. (1972): "Regression models and life-tables," *Journal of the Royal Statistical Society. Series B (Methodological)*, 34, 187–220.

Dabrowska, D. (1987): "Non-parametric regression with censored survival time data," *Scandinavian Journal of Statistics*, 14, 181–197.

Dabrowska, D. (1989): "Uniform consistency of the kernel conditional kaplan-meier estimate," *The Annals of Statistics*, 17, 1157–1167.

Dette, H. and S. Volgushev (2008): "Non-crossing non-parametric estimates of quantile curves," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70, 609–627.

Doksum, K. A. and M. Gasko (1990): "On a correspondence between models in binary regression analysis and in survival analysis," *International Statistical Review*, 58, 243–252.

Efron, B. (1981): "Censored data and the bootstrap," *Journal of the American Statistical Association*, 76, 312–319.

Fine, J. P. (2001): "Regression modeling of competing crude failure probabilities," *Biostatistics*, 2, 85–97.

Gerds, T. and M. Schumacher (2006): "Consistent estimation of the expected Brier score in general survival models with right-censored event times," *Biometrical Journal*, 48, 1029–1040.

Gerds, T. A. (2013): "prodlim: Product limit estimation for event history and survival analysis," http://CRAN.R-project.org/package=prodlim, R package version 1.3.7. Available from: http://CRAN.R-project.org/package=prodlim.

Gneiting, T. and A. Raftery (2007): "Strictly proper scoring rules, prediction, and estimation," *Journal of the American Statistical Association*, 102, 359–378.

González Manteiga, W. and C. Cadarso-Suarez (1994): "Asymptotic properties of a generalized kaplan-meier estimator with some applications," *Journal of Nonparametric Statistics*, 4, 65–78.

Graf, E., C. Schmoor, W. Sauerbrei, and M. Schumacher (1999): "Assessment and comparison of prognostic classification schemes for survival data," *Statistics in Medicine*, 18, 2529–2545.

Grambsch, P. and T. Therneau (1994): "Proportional hazards tests and diagnostics based on weighted residuals," *Biometrika*, 81, 515–526.

Hasford, J., M. Pfirrmann, R. Hehlmann, and Others (1998): "A new prognostic score for survival of patients with chronic myeloid leukemia treated with interferon alfa," *Journal of the National Cancer Institute*, 90, 850–858.

Hehlmann, R., H. Heimpel, J. Hasford, and Others (1994): "Randomized comparison of interferon-α with busulfan and hydroxyurea in chronic myelogenous leukemia," *Blood*, 84, 4064–4077.

Herberich, E. and T. Hothorn (2012): "Dunnett-type inference in the frailty Cox model with covariates," *Statistics in Medicine*, 31, 45–55.

Honoré, B., S. Khan, and J. Powell (2002): "Quantile regression under random censoring," *Journal of Econometrics*, 109, 67–105.

Hothorn, T. (2013): "ctmDevel: Conditional transformation models," https://r-forge.r-project.org/projects/ctm, R package version 0.1-0. SVN revision 56. Available from: https://r-forge.r-project.org/projects/ctm.

Hothorn, T., P. Bühlmann, S. Dudoit, A. Molinaro, and M. J. van der Laan (2006a): "Survival ensembles," *Biostatistics*, 7, 355–373.

Hothorn, T., K. Hornik, and A. Zeileis (2006b): "Unbiased recursive partitioning: A conditional inference framework," *Journal of Computational and Graphical Statistics*, 15, 651–674.

Hothorn, T., T. Kneib, and P. Bühlmann (2014): "Conditional transformation models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76, 3–27.

Hothorn, T., B. Lausen, A. Benner, and M. Radespiel-Tröger (2004): "Bagging survival trees," *Statistics in Medicine*, 23, 77–91.

Iglesias Pérez, C. and W. González Manteiga (1999): "Strong representation of a generalized product-limit estimator for truncated and censored data with some applications," *Journal of Nonparametric Statistics*, 10, 213–244.

Ishwaran, H., U. B. Kogalur, E. H. Blackstone, and M. S. Lauer (2008): "Random survival forests," *The Annals of Applied Statistics*, 2, 841–860.

Kaplan, E. and P. Meier (1958): "Nonparametric estimation from incomplete observations," *Journal of the American Statistical Association*, 53, 457–481.

Van der Laan, M. J. and J. M. Robins (2003): *Unified Methods for Censored Longitudinal Data and Causality*, New York: Springer.

Lee, K. H., S. Chakraborty, and J. Sun (2011): "Bayesian variable selection in semiparametric proportional hazards model for high dimensional survival data," *The International Journal of Biostatistics*, 7, Article 21.

Li, G. and H. Doss (1995): "An approach to nonparametric regression for life history data using local linear fitting," *The Annals of Statistics*, 23, 787–823.

Lin, D. Y., L. J. Wei, and Z. Ying (1993): "Checking the Cox model with cumulative sums of martingale-based residuals," *Biometrika*, 80, 557–572.

Lu, W. and L. Li (2008): "Boosting method for nonlinear transformation models with censored survival data," *Biostatistics*, 9, 658–667.

Mackenzie, T. (2012): "Survival curve estimation with dependent left truncated data using Cox's model," *The International Journal of Biostatistics*, 8, Article 29.

Mackillop, W. J. and C. F. Quirt (1997): "Measuring the accuracy of prognostic judgments in oncology," *Journal of Clinical Epidemiology*, 50, 21–29.

McGilchrist, C. and C. Aisbett (1991a): "Regression with frailty in survival analysis," *Biometrics*, 47, 461–466.

McGilchrist, C. A. and C. W. Aisbett (1991b): "Regression with frailty in survival analysis," *Biometrics*, 47, 461–466.

McKeague, I. and K. Utikal (1990): "Inference for a nonlinear counting process regression model," *The Annals of Statistics*, 18, 1172–1187.

Meinshausen, N. (2006): "Quantile regression forests," *Journal of Machine Learning Research*, 7, 983–999.

Mogensen, U. B., H. Ishwaran, and T. A. Gerds (2012): "Evaluating random forests for survival analysis using prediction error curves," *Journal of Statistical Software*, 50, 1–23.

Möst, L., M. Schmid, F. Faschingbauer, and T. Hothorn (2014): "Predicting birth weight with conditionally linear transformation models," *Statistical Methods in Medical Research*.

Ng'Andu, N. H. (1997): "An empirical comparison of statistical tests for assessing the proportional hazards assumption of Cox's model," *Statistics in Medicine*, 16, 611–626.

Peng, L. and Y. Huang (2008): "Survival analysis with quantile regression models," *Journal of the American Statistical Association*, 103, 637–649.

Portnoy, S. (2003): "Censored regression quantiles," *Journal of the American Statistical Association*, 98, 1001–1012.

Powell, J. (1986): "Censored regression quantiles," *Journal of Econometrics*, 32, 143–155.

R Core Team (2013): "R: A language and environment for statistical computing," http://www.R-project.org/, available from: http://www.R-project.org/.

Robins, J. M. and D. M. Finkelstein (2000): "Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests," *Biometrics*, 56, 779–788.

Sargent, D. (1997): "A flexible approach to time-varying coefficients in the cox regression setting," *Lifetime Data Analysis*, 3, 13–25.

Scheike, T. and T. Martinussen (2004): "On estimation and tests of time-varying effects in the proportional hazards model," *Scandinavian Journal of Statistics*, 31, 51–62.

Schemper, M. (1992): "Cox analysis of survival data with non-proportional hazard functions," *Journal of the Royal Statistical Society. Series D (The Statistician)*, 41, 455–465.

Schemper, M. and R. Henderson (2000): "Predictive accuracy and explained variation in Cox regression," *Biometrics*, 56, 249–255.

Schmid, M. and T. Hothorn (2008): "Boosting additive models using component-wise P-splines as base-learners," *Computational Statistics & Data Analysis*, 53, 298–311.

Schoenfeld, D. (1982): "Partial residuals for the proportional hazards regression model," *Biometrika*, 69, 239–241.

Spierdijk, L. (2008): "Nonparametric conditional hazard rate estimation: A local linear approach," *Computational Statistics & Data Analysis*, 52, 2419–2434.

Strobl, C., A.-L. Boulesteix, A. Zeileis, and T. Hothorn (2007): "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC Bioinformatics*, 8, 25.

Therneau, T. M. (2013): "Survival analysis," http://CRAN.R-project.org/package=survival, R package version 2.37-4. Available from: http://CRAN.R-project.org/package=survival.

Tian, L., D. Zucker, and L. Wei (2005): "On the Cox model with time-varying regression coefficients," *Journal of the American Statistical Association*, 100, 172–183.

Van der Vaart, A. and M. J. van der Laan (2006): "Estimating a survival distribution with current status data and high-dimensional covariates," *The International Journal of Biostatistics*, 2, Article 9.

Vaida, F. and R. Xu (2000): "Proportional hazards model with random effects," *Statistics in Medicine*, 19, 3309–3324.

Wang, H. and L. Wang (2009): "Locally weighted censored quantile regression," *Journal of the American Statistical Association*, 104, 1117–1128.

Wey, A., L. Wang, and K. Rudser (2014): "Censored quantile regression with recursive partitioning-based weights," *Biostatistics*, 15, 170–181.

Xu, R. and J. O'Quigley (2000): "Estimating average regression effect under non-proportional hazards," *Biostatistics*, 1, 423–439.

Zeng, D. and D. Y. Lin (2006): "Efficient estimation of semiparametric transformation models for counting processes," *Biometrika*, 93, 627–640.